Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 135-138 recibido 16-03-2018 revisado 29-04-2018 aceptado 03-05-2018

ARAP: Arabic Author Profiling Project for Cyber-Security

ARAP: Proyecto sobre Perfiles de Autoría en Árabe para la Ciber-Seguridad

Paolo Rosso[1], **Francisco Rangel**[1], **Bilal Ghanem**[1], **Anis Charfi**[2]
[1]PRHLT Rsearch Center, Universitat Politècnica de València
Camino de Vera s/n, 46022 Valencia, Spain
prosso@dsic.upv.es, francisco.rangel@autoritas.es, bigha@doctor.upv.es
[2]Cargenie Mellon University Qatar, Education City PO Box 24866 Doha, Qatar
acharfi@andrew.cmu.edu

Abstract: In this paper we describe the current state of the ARAP project on Arabic Author Profiling for Cyber-Security funded by the Qatar National Research Fund via the Carnegie Mellon University in Qatar. The project focuses on determining whether a suspicious message is actually a potential threat and, in that case, aims at profiling its author. The contribution of the project lies in the lack of research of this type for Arabic.
Keywords: Arabic, cyber-security, author profiling, gender, age, native language, language variety, deception, irony

Resumen: En este artículo describimos el estado actual del proyecto ARAP de perfilado de autores en árabe para la ciberseguridad financiado por la Qatar National Research Fund va la Universidad de Carnegie Mellon en Qatar. El proyecto se centra en determinar cuando un mensaje sospechoso realmente es una amenaza potencial, y en tal caso, perfilar a su autor. La contribución del proyecto reside en la falta de investigaciones de este tipo para el árabe.
Palabras clave: Árabe, ciberseguridad, perfiles de autor, sexo, edad, idioma nativo, variedad del lenguaje, engaño, ironía

1 Cyber-Security in Social Media

The anonymity of social media provides with new ways of communication without censorship. However, the lack of knowledge about the authors may contribute to new cyber-security issues, such as threatening messages or terrorism propaganda. Security in the fifth domain, the cyber space, is nowadays one of the defense priorities in many nations[1]. Generating intelligence from social networks content is important to prevent cyber threats. To profile potential terrorists from the messages that they share in their social circles allows detecting communities whose aim is to undermine the security in our daily life. From 2014 to 2017, the PRHLT research center of the Universitat Politècnica de València has been involved in a research project funded by the Army Research Office of the United States whose objective was the detection of communities in Twitter that shared content about ISIS (large-scale copy detection in social circles)[2]. Since 2017 the PRHLT research center takes part in a project of the Qatar National Research Fund whose aim is determining the linguistic profile of the author of a suspicious or threatening text, with the attempt of profiling potential terrorists (Russell and Miller, 1977). When a suspicious message is analysed, we follow the workflow in Figure 1. Concretely: *(i)* we check the veracity of the threat, discarding those messages that are deceptive (Cagnina and Rosso, 2017) or ironic (Hernández, Patti, and Rosso, 2016), since they do not represent a real threat; finally, *(ii)* we profile the demographics of its author (Rangel and Rosso, 2016) as well as

[1]https://www.economist.com/node/16478792

[2]https://www.prhlt.upv.es/wp/es/project/2016/sococode

ISSN 1135-5948. DOI 10.26342/2018-61-16

her cultural and social context.

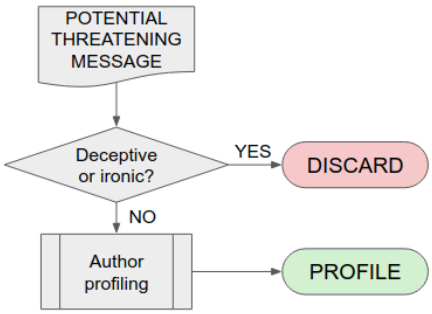

Figure 1: Workflow to profile the author of a potential threatening message.

The contribution of the presented project is relevant due to the lack of this kind of investigations in the Arabic language (Rosso et al., 2018). In the next sections, we describe the current state of the project.

2 Threat Messages: Beyond Deceptive Messages and Irony

A suspicious message may not be a threat when it is deceptive or ironic. A message can be considered deceptive when it is written with the intention to sound authentic. Deceptive detection has mainly focused on the detection of spam in opinion reviews (Cagnina and Rosso, 2017) and nowadays there is an increasingly interest in fake news detection (e.g. deception in political debates[3]), also in Arabic. Despite the interest in the area, few are the preliminary works in Arabic in deception detection in reviews and news.

The most common definition of irony is the use of words to express the opposite meaning from what is said. With respect to irony detection in Arabic, even less are the preliminary works. Recently, a preliminary system for irony detection in Arabic in social media was presented in (Karoui, Zitoune, and Moriceau, 2017). At the moment we are helping in the annotation of an enriched version of the corpus the authors employed in the previous work.

3 Author Profiling: Gender, Age, Native Language, and Arabic Language Variety

Since 2013 author profiling in social media has been addressed in the framework of the PAN Lab[4] at CLEF[5]. We started addressing the Arabic language since 2017 (Rangel et al., 2017): gender and language variety identification.

In the framework of the ARAP project, we approached both with the Low Dimensionality Statistical Embedding (LDSE) (Rangel, Rosso, and Franco-Salvador, 2018) representation. This method represents documents on the basis of the probability distribution of occurrence of their words in the different classes. The key concept of LDSE is a weight, representing the probability of a term to belong to one of the different categories: for gender (female vs. male) and for language variety (e.g. Egypt vs. Levantine vs. ...). The distribution of weights for a given document should be closer to the weights of its corresponding category. In order to compare the obtained results, we have proposed the following two state-of-the-art baselines: *i)* BASELINE-stat that emulates random choice, depending on the number of classes. For example, in gender identification with two balanced classes, this baseline obtain 50% of accuracy; and *ii)* BASELINE-bow that represents documents as a bag-of-words with the 1,000 most common words in the training set, weighted by absolute frequency of occurrence. The texts are preprocessed as follows: lowercase words, removal of punctuation signs and numbers, and removal of stop words. A Support Vector Machine classifier with a linear kernel and default parameters is used.

In the next subsections, we describe the corpora used and the obtained results with LDSE compared to the described baselines. In case of using the PAN corpus, we have also compared ourselves with the best result obtained in the official task.

3.1 Gender Identification

We have collected two corpora and annotated with gender information: *i)* PAN-AP-2017[6]; and *ii)* CMUQ-ARAP. To build the PAN-AP-2017 corpus, we have retrieved tweets geolocated in the following cities: Cairo, Abu Dhabi, Doha, Kuwait, Manama, Mascate, Riyadh, Sana'a, Amman, Beirut, Damascus,

[3] http://alt.qcri.org/clef2018-factcheck

[4] http://pan.webis.de

[5] http://clef2018.clef-initiative.eu

[6] The PAN-AP-2017 corpus has been also annotated with language variety information. Hence, the methodology described here applies to the next section.

Jerusalem, Algiers, Rabat, Tripoli, and Tunis. We have selected the unique authors who wrote these tweets and downloaded their timelines. Authors with their location outside the given regions have been discarded, as well as retweets or tweets written in other languages than Arabic. We have annotated gender automatically with a dictionary of proper nouns and performed a manual review of their profiles to fix errors and discard ambiguous profiles. The corpus consists of a total of 4,000 authors completely balanced by gender, with 100 tweets per author, and split into training and test following 60/40 proportion. The Carnegie Mellon University in Qatar has developed the CMUQ-ARAP corpus. This corpus consists of a total of 10,140 authors with a variable number of tweets (from a few ones to thousands), imbalanced with respect to gender, and split into training and test following 60/40 proportion. Corpora statistics are shown in Table 1.

Set	Males	Females	Total
PAN-AP-2017			
Training	1,200	1,200	2,400
Test	800	800	1,600
Total	2,000	2,000	4,000
CMUQ-ARAP			
Training	5,481	3,645	9,126
Test	609	405	1,014
Total	6,090	4,050	10,140

Table 1: Arabic corpora annotated with gender information.

The following tweets are examples written by a female and a male respectively:

هههه كنت مشغولة يومها عندي بزنس مش فاضية

انا مش متخيل حيجي يوم نسوق في كوبري

In the female sentence (*hahaha I was busy on that day, I have a business so I'm not free*), the words مشغولة (*busy*) and فاضية (*free*) indicate that the writer is female. In the Arabic language when the adjective ends with ـة (*Taa': one of the Arabic letter*) (this is the letter shape in an independent situation: ة) letter, it gives an indicator that the speaker is a female. While in the male sentence (*I don't imagine that in the future we will drive in a subway*), the word متخيل (*conceived*) without the same previous letter (ة) implies that the speaker is a man.

We have used the LDSE representation with SVM with Gaussian kernel and gamma equal to 0.02 to approach the task on the PAN-AP-2017 corpus, whereas SVM with linear kernel for the CMUQ-ARAP corpus[7]. The obtained results in terms of accuracy can be seen in Table 2.

Corpus	Method	Accuracy
PAN-AP-2017	LDSE	73.19
	BASELINE-bow	53.00
	BASELINE-stat	50.00
	Best at PAN'17	80.31
CMUQ-ARAP	LDSE	66.96
	BASELINE-bow	61.56
	BASLINE-stat	60.06

Table 2: Gender results in terms of accuracy.

In case of PAN-AP-2017 corpus, the proposed method obtains about 20% higher accuracy than the best baseline (38% of improvement), although it obtains lower results than the best participant at PAN (7.12%). In case of CMUQ-ARAP, the proposed method obtains about 5.4% higher accuracy than the best baseline (8.77% of improvement). Despite the imbalance in the number of authors per gender besides the uneven number of tweets per author, the results are not much lower than the obtained with the PAN-AP-2017 corpus[8].

3.2 Language Variety Identification

Following previous works (Sadat, Kazemi, and Farzindar, 2014), the aforementioned PAN-AP-2017 corpus has been annotated with four varieties of Arabic: Egypt, Gulf, Levantine and Maghrebi. There are 1,000 authors per variety, divided into 600/400 for training and test respectively. Each author contains 100 tweets.

The following tweet is an example of the Gulf language variety:

طاريك بمليني فرح ، شعاد لقياك

(*Your remembrance makes me rejoice, but what about a meeting with you!*). The words طاريك (*your remembrance*), شعاد (*but what*

[7]Several algorithms and parameters have been tested and we have selected the configuration that obtained the best results.

[8]The CMUQ team is working on balancing the corpus in terms of number of authors per gender and number of tweets per author.

about) and لقياك (*meeting you*) are only used in the Gulf variety.

Table 3 shows the results for the language variety task on the PAN-AP-2017 corpus. We have used LDSE and SVM with Gaussian kernel and default parameters. As can be seen, LDSE outperforms both baselines in 48.56% and 57.50% of accuracy (143% and 230% of improvement respectively). The difference with respect to the best result at PAN (0.63%) is not statistically significant.

Method	Accuracy
LDSE	82.50
BASELINE-bow	33.94
BASELINE-stat	25.00
Best at PAN'17	83.13

Table 3: Gender results in terms of accuracy.

4 Conclusions and Future Work

In this paper we have presented the current state of the ARAP project on Arabic Author Profiling for Cyber-Security. So far, we have focused mainly on gender and language variety identification since 2017. We have included the Arabic language in the organisation of the PAN shared task and built corpora labeled with gender and language variety information. We have proposed a method to approach both problems and obtained competitive results. At the moment of writing of this paper, the colleagues at the Carnegie Mellon University in Qatar are tagging the CMUQ-ARAP corpus with the information about age. Moreover, the colleague in Qatar are organising the Fact Checking Lab[9] at CLEF on deceptive detection in political debates in English and Arabic.

As future work, at PAN@CLEF in 2018 we will address gender identification in Twitter from a multimodal perspective taking into account not only the textual information but also the images of the URLs links in tweets. In the future, together with the Carnegie Mellon University in Qatar, we plan to organise a track at the Forum of the Information Retrieval Evaluation addressing the several aspects of the ARAP research project.

Agradecimientos

This article was made possible by NPRP grant 9-175-1-033 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

[9] http://alt.qcri.org/clef2018-factcheck

References

Cagnina, L. C. and P. Rosso. 2017. Detecting deceptive opinions: Intra and cross-domain classification using an efficient representation. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 25(Suppl. 2):151–174.

Hernández, I., V. Patti, and P. Rosso. 2016. Irony detection in twitter: The role of affective content. *ACM Transactions on Internet Technology (TOIT)*, 16(3):19.

Karoui, J., F. B. Zitoune, and V. Moriceau. 2017. Soukhria: Towards an irony detection system for arabic in social media. *Procedia Computer Science*, 117:161–168.

Rangel, F. and P. Rosso. 2016. On the impact of emotions on author profiling. *Information processing & management*, 52(1):73–92.

Rangel, F., P. Rosso, and M. Franco-Salvador. 2018. A low dimensionality representation for language variety identification. In *CICLing-2016, Springer-Verlag, Revised Selected Papers, Part II, LNCS(9624)*. Springer-Verlag, arXiv:1705.10754.

Rangel, F., P. Rosso, M. Potthast, and B. Stein. 2017. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. In *CLEF 2017 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1866.*

Rosso, P., F. Rangel, I. Hernández, L. Cagnina, W. Zaghouani, and A. Charfi. 2018. A survey on author profiling, deception, and irony detection for the arabic language. *Language and Linguistics Compass (in press).*

Russell, C. A. and B. H. Miller. 1977. Profile of a terrorist. *Studies in Conflict & Terrorism*, 1(1):17–34.

Sadat, F., F. Kazemi, and A. Farzindar. 2014. Automatic identification of arabic language varieties and dialects in social media. *Proceedings of SocialNLP*, page 22.